

# Finite-Sample Equivalence of Several Statistical Models for Presence-Only Data

William Fithian  
Department of Statistics  
Stanford University  
wfithian@stanford.edu

Trevor Hastie  
Department of Statistics  
Stanford University  
hastie@stanford.edu

July 31, 2012

## Abstract

Statistical modeling of presence-only data has attracted much recent attention in the ecological literature, leading to a proliferation of methods, including the inhomogeneous poisson process (IPP) model [12], maximum entropy (Maxent) modeling of species distributions [8] [5] [6], and logistic regression models. Several recent articles have shown the close relationships between these methods [1] [12]. We explain why the IPP intensity function is a more natural object of inference in presence-only studies than occurrence probabilities. Further, we explain why the above three techniques all essentially implement the IPP model, and prove exact finite-sample equivalence between the IPP model, Maxent, and a weighted form of logistic regression. That is, given the same data, the same basis expansions, and the same regularization penalty — and provided that the intercept  $\alpha$  is unpenalized — the slope coefficient estimates  $\hat{\beta}$  are identical for each method. The exact equivalence of these various models implies that every method is equally well-motivated and equally extensible with existing software packages. In particular, many methods that extend logistic regression can also extend Maxent or the IPP model in exactly analogous ways.

At several points we discuss issues of imperfect detection and observer bias, and describe how to use the IPP model to combine presence-only and presence-absence records from one or more species to address these issues.

## 1 Introduction

A common ecological problem is estimating where a species of interest can be found from records of where it has been found in the past. There are many motivations for solving this problem: planning wildlife management actions, monitoring endangered or invasive species, or simple scientific discovery.

### 1.1 Presence-Only Data

Estimation is simplest and most convincing when the observations of species presence are collected in a systematic manner. In a typical design, a surveyor

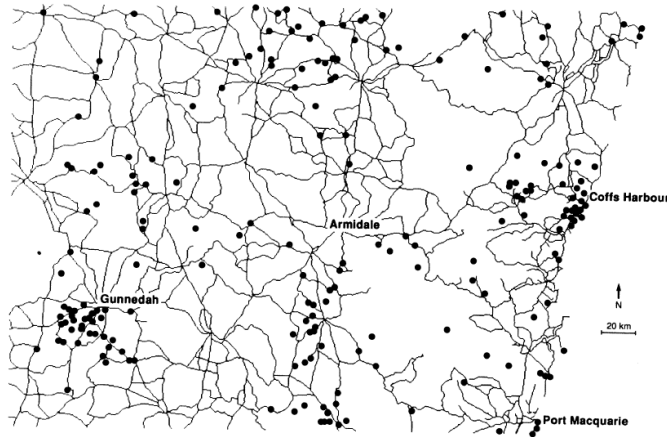


Figure 1. Koala records (courtesy of New South Wales National Parks & Wildlife Service) and the road network on part of the New South Wales north coast.

Figure 1: Observer bias in presence-only data for koalas. Taken from Margules and Austen (1994).

visits a one-hectare patch of land for one hour and records whether or not she discovers any specimens in that interval. The records of unsuccessful surveys are called absence records — a mild misnomer since ecologists recognize that specimens could be present but go undetected. Data sets of this form are called presence-absence data.

Unfortunately, presence-absence data are often expensive to collect, especially for rare or elusive species. For many species of interest, the only data available are museum or herbarium records of locations where a specimen was found. These presence-only records are often collected haphazardly and frequently suffer from unknown observer bias such as that illustrated in Figure 1. The clustering of koala sightings near roads and cities probably has more to do with the behavior of people than of koalas.

## 1.2 What Should We Estimate?

Before we can sensibly decide how to model presence-only data, we must address the issue of what it is we are modeling in the first place. How should we even think of “occurrence,” the scientific phenomenon nominally under study? This issue arises with presence-only and presence-absence data alike.

### 1.2.1 Occurrence Probability

To illustrate this conundrum, we include a typical “heat-map” output (Figure 2) of a study of the willow tit in Switzerland using count data [9]. The map reveals which locations are favored by the species and which are not (in this case, high elevation and moderate forest cover appear to be the bird’s habitat of choice). The legend shows that the color of a region reflects the local probability of “occurrence.”

But precisely what event has this probability? Reading the paper, we discover that occurrence means that there is at least one willow tit present on a

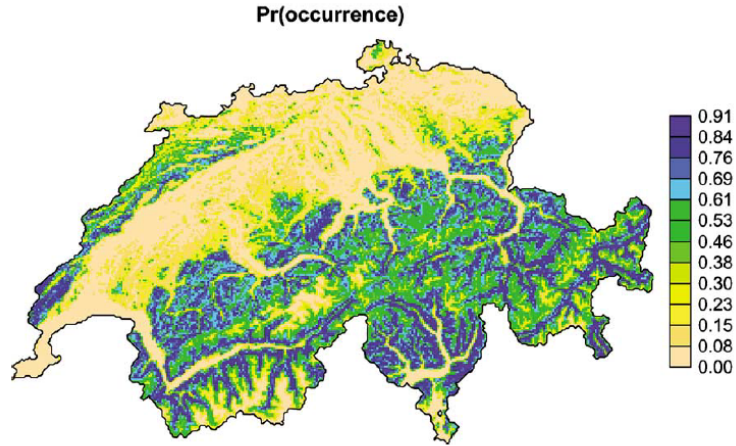


Figure 2: Typical heat map of occurrence probabilities. Taken from Royle et al. (2005)

survey path through a  $1 \text{ km} \times 1 \text{ km}$  quadrat of land. The authors analyze a presence-absence data set using a hierarchical model that explicitly accounts for the possibility that a bird was present but not detected at the time of the survey.

Because the survey path length varies across sampling units, the authors use it in their model as a predictor of presence probability. It is not specified which value of this predictor is used in generating the heat map, which makes the map difficult to interpret.

Even if we could interpret the heat map as the probability of a bird being present anywhere in the quadrat (not just along the path), the probability of a bird being present would still be larger in a  $2 \text{ km} \times 2 \text{ km}$  sampling unit and smaller in a  $100 \text{ m} \times 100 \text{ m}$  one. Therefore the very definition of “occurrence probability” in a presence-absence study depends on the specific sampling scheme used to collect the presence-absence data. Correspondingly, interpreting the legend of such a heat map can only make sense in the context of a specific quadrat size (typically, whatever size was used in the study).

Though the choice to estimate occurrence probability in a specific quadrat size is ecologically arbitrary, it can at least yield estimates with a meaningful interpretation. By contrast, trying to estimate “occurrence probability” in a presence-only study is a far murkier proposition. Any method that estimates occurrence probabilities without reference to quadrat size would seem to be predicting the same probability of occurrence within a large quadrat or a small one, which cannot make sense.

### 1.2.2 Occurrence Rate

Since occurrence probability is only meaningful with reference to a specific quadrat size, it is an awkward quantity to model in a presence-only study. We feel that it is more natural to estimate an occurrence rate in this context, a quantity with units of inverse area (e.g.  $1/\text{km}^2$ ) corresponding to the expected

number of specimens per unit area. Under some simple stochastic models for species occurrence, including the ones considered here, specifying the occurrence rate is equivalent to specifying occurrence probability simultaneously for all quadrat sizes.

Unfortunately, estimating this rate as an absolute quantity is not practical for typical presence-only samples. For instance, suppose our koala data set had twice as many sightings. It might mean that there were twice as many koalas, or it might only mean that there were the same number of koalas but twice as many motorists to see them, or that motorists were twice as likely to call in a sighting given that it occurred.

Since the total number of sightings reflects nothing meaningful about the occurrence rate, the only real information at our disposal is the distribution of sightings across our domain of interest. If we are lucky, this distribution reflects the *relative* occurrence rate of the species, i.e. a function proportional to the occurrence rate. Even trying to estimate relative occurrence rates is susceptible to confounding as a result of observer bias, an issue we explore in more detail in Section 2.5.

### 1.3 Notation

We now introduce notation we will use for the remainder of the article. We begin with some geographic domain of interest  $\mathcal{D}$ , typically a bounded subset of  $\mathbb{R}^2$ . If the time of an observation is an important variable, we might take  $\mathcal{D} \subseteq \mathbb{R}^3$ , giving a point process in both space and time.

Assume without loss of generality that the total area of  $\mathcal{D}$  is 1. Associated with each geographic location  $z \in \mathcal{D}$  is a vector  $x(z)$  of measured features.

Our presence-only data set consists of  $n_1$  locations  $z_i \in \mathcal{D}$ ,  $i = 1, 2, \dots, n_1$ . This data set is accompanied by  $n_0$  “background” observations  $z_i$ ,  $i = n_1 + 1, \dots, n_1 + n_0$ , typically a simple random sample from  $\mathcal{D}$ .

Finally let  $x_i = x(z_i)$  be the features associated with observation  $i$ , and  $y_i = \mathbf{1}_{i \leq n_1}$  be an indicator of presence/background status.

Our treatment of these data as random or fixed will vary throughout the article.

### 1.4 Outline

The rest of the paper is organized as follows. In Section 2 we define the log-linear inhomogeneous poisson process (IPP) model and its application to presence-only data, with special focus on interpretations of the model and its score equations. Aarts et al (2011) showed that many methods in species modeling can be motivated by the IPP model. We explicitly draw these connections for several especially illuminating examples.

In Section 3 we consider a particularly important example, showing that the Maxent likelihood of Phillips et al. follows immediately from partially maximizing the IPP log-likelihood with respect to the intercept  $\alpha$ . It is equivalent to the IPP in the sense that both methods produce exactly the same slope estimates  $\hat{\beta}$  in any finite sample, and the intercept estimate  $\hat{\alpha}$  from an IPP is rarely scientifically interesting.

In Section 4 we discuss logistic regression and its connections to the IPP model. Warton and Shepherd (2010) showed that the logistic regression can

be derived from the IPP model, and that for any fixed presence sample as the number of background samples tends to infinity the fitted logistic regression slopes converge to the fitted IPP slopes.

However, if the log-linear model is misspecified this convergence may not occur until the background sample is extremely large, and may not occur at all if the number of presence samples grows along with the number of background samples.

We show that by reweighting the samples we can use logistic regression to recover the IPP estimate  $\hat{\beta}$  precisely in any finite sample.

An advantage of having IPP as a unifying model for presence-only data is that we can derive a variety of other simple parametric forms for other types of data, including presence-absence and count data. In Section 5 we consider how we might use this fact to join disparate data sets into one likelihood function, either to share information across species, to estimate an overall abundance rate, or even to estimate the level of observer bias in presence-only samples. Section 6 contains discussion.

## 2 The Inhomogeneous Poisson Process Model

The IPP is a simple model for the distribution of a random set of points  $\mathbf{Z}$  falling in some domain  $\mathcal{D}$ . Both the number of points and their locations are random.

An IPP can be defined by its intensity function

$$\lambda : \mathcal{D} \longrightarrow [0, \infty) \quad (1)$$

Informally,  $\lambda$  indexes the likelihood that a point falls at or near  $z$ . For subsets  $A \subseteq \mathcal{D}$ , define

$$\Lambda(A) = \int_A \lambda(z) dz \quad (2)$$

and assume  $\Lambda(\mathcal{D}) < \infty$ .

There are two main ways to formally characterize an IPP with intensity  $\lambda$ . One simple definition is that the total number of points is a Poisson random variable with mean  $\Lambda(\mathcal{D})$ , and their locations are independent and identically distributed with density  $p_\lambda(z) = \lambda(z)/\Lambda(\mathcal{D})$ . The only difference between an IPP and a simple random sample from  $p_\lambda$  is that its size is random.

Alternatively, we can think of an IPP as a continuous limit of a poisson count model in discretized geometric space. Let  $N(A) = \#(\mathbf{Z} \cap A)$ , the number of points falling in set  $A$ . An equivalent characterization of the IPP model is that

$$N(A) \sim \text{Poisson}(\Lambda(A)) \quad (3)$$

with  $N(A)$  and  $N(B)$  independent for disjoint sets  $A$  and  $B$ . For more on the IPP and other point process models, see e.g. [2].

In the case of a finite discrete domain  $\mathcal{D} = \{z_1, z_2, \dots, z_m\}$ , the IPP model reduces to a discrete Poisson model, with

$$N(z_i) \sim \text{Poisson}(\lambda(z_i)) \quad (4)$$

In this sense, the IPP model is a limit of finer and finer discretizations of  $\mathcal{D}$ .

## 2.1 Modeling Presence-Only Data as an IPP

Warton and Shepherd (2010) proposed modeling the species sightings  $z_1, \dots, z_{n_1}$  as arising from an IPP sightings process whose intensity is a log-linear function of the features  $x(z)$ :

$$\lambda(z) = e^{\alpha + \beta' x(z)} \quad (5)$$

The linearity assumption does not impose as many restrictions as it may appear to, since our choice of feature vector  $x(z)$  could incorporate polynomial terms, interactions, a spline basis, or other basis expansions.

If we adopt the interpretation of an IPP as a simple random sample with random size, we see that  $\alpha$  and  $\beta$  play very different roles. Since  $\alpha$  only scales  $\lambda(z)$  up or down, it has no effect on  $p_\lambda = \lambda/\Lambda(D)$ . The “slope” parameters  $\beta$  completely determine  $p_\lambda$ , while  $\alpha$  merely scales the intensity up or down to attain any desired average sample size  $\Lambda(D)$ .

## 2.2 Maximum Likelihood for the IPP

Like many exponential family models, the log-linear IPP has simple and enlightening score equations. The log-likelihood is

$$\ell(\alpha, \beta) = \sum_{y_i=1} \alpha + \beta' x_i - \int_{\mathcal{D}} e^{\alpha + \beta' x(z)} dz \quad (6)$$

Differentiating with respect to  $\alpha$  we obtain the score equation

$$n_1 = \int_{\mathcal{D}} e^{\alpha + \beta' x(z)} dz = \Lambda(D) \quad (7)$$

That is, whatever  $\hat{\beta}$  is,  $\hat{\alpha}$  plays the role of a “normalizing” constant guaranteeing that  $\lambda(z)$  integrates to  $n_1$ . This is our first glimpse at why  $\hat{\alpha}$  is typically not of scientific interest, since it basically encodes the total number of museum records we have.

Solving for  $\alpha$  in (7) we obtain the partially maximized log-likelihood

$$\ell^*(\beta) = \sum_{y_i=1} \left( \log n_1 - \log \left( \int_{\mathcal{D}} e^{\beta' x(z)} dz \right) + \beta' x_i \right) - n_1 \quad (8)$$

Rearranging terms and ignoring constants, we have

$$\ell^*(\beta) = \sum_{y_i=1} \beta' x_i - n_1 \log \left( \int_{\mathcal{D}} e^{\beta' x(z)} dz \right) \quad (9)$$

$$= \sum_{y_i=1} \log p_\lambda(z_i) \quad (10)$$

the same log-likelihood we would obtain by conditioning on  $n_1$  and treating  $z_i$  as a simple random sample from  $p_\lambda$ .

Differentiating (9) with respect to  $\beta$  and dividing by  $n_1$  gives the remaining score equations:

$$\frac{1}{n_1} \sum_{y_i=1} x_i = \frac{\int_{\mathcal{D}} e^{\beta' x(z)} x(z) dz}{\int_{\mathcal{D}} e^{\beta' x(z)} dz} \quad (11)$$

$$= \mathbb{E}_{p_\lambda} x(z) \quad (12)$$

This amounts to finding  $\beta$  for which the expectation of  $x(z)$  under  $p_\lambda(z)$  matches the empirical mean of the presence samples.

Maximizing the likelihood of a log-linear IPP model, then, amounts to

1. Choosing  $\beta$  so  $p_\lambda$  matches the means of the features  $x(z)$  in the presence sample.
2. Choosing  $\alpha$  so that  $\lambda(z) = n_1 p_\lambda(z)$ .

The first step is really a parametric density estimation problem for the presence sample, and the second step doesn't matter if  $n_1$  has no scientific interpretation. Whenever we don't care about  $n_1$ , the IPP is really nothing more than a density estimation procedure (and only a little more complicated if we do).

Despite our denigration of  $n_1$  (and therefore  $\hat{\alpha}$ ) as a quantity of scientific interest, there is one exception we can think of. When several species are under consideration, it might be interesting that species 1 was sighted twice as often as species 2 — especially if we can obtain an independent estimate of the true abundance level of species 2, say through presence-absence data. We expand upon this idea in Section 5.

### 2.3 Numerical Evaluation of the Integral

The IPP likelihood and score equations involve integrals that, in general, we cannot evaluate analytically. However, we can use the background samples to evaluate it via Monte Carlo. Since  $\int_{\mathcal{D}} dz = 1$ , we can approximate any  $\int_{\mathcal{D}} f(z) dz$  by an average  $m^{-1} \sum_{i=1}^m f(z_i)$  over a uniform random sample  $z_i$ . If our background points comprise such a sample, we can replacing the original log-likelihood with

$$\ell(\alpha, \beta) = \sum_{y_i=1} \alpha + \beta' x_i - \frac{1}{n_0} \sum_{y_i=0} e^{\alpha + \beta' x_i} \quad (13)$$

Two other options for numerically evaluating the integral are to choose background points in a regular fine grid of  $\mathcal{D}$ , or to assign quadrature weights to the background points and approximate the integral with a weighted sum. In the first case, the optimization criterion would be the same, and in the second the only difference would be that the second sum would be a weighted sum over the background points.

Repeating the previous derivation gives the numerical version of the score equations

$$n_1 = \frac{1}{n_0} \sum_{y_i=0} e^{\alpha + \beta' x_i} \quad (14)$$

$$\frac{1}{n_1} \sum_{y_i=1} x_i = \frac{n_0^{-1} \sum_{y_i=0} e^{\beta' x_i} x_i}{n_0^{-1} \sum_{y_i=0} e^{\beta' x_i}} \quad (15)$$

Throughout, we will refer to (13) as the numerical IPP log-likelihood to distinguish it from the true IPP log-likelihood (6). In practice, “fitting” the IPP model would mean fitting (13).

## 2.4 Connection to Poisson Log-Linear Model

Suppose that  $x(z)$  is a continuous function on  $\mathcal{D}$ . If we use background points from a regular fine grid, we are essentially discretizing  $\mathcal{D}$  into  $n_0$  pixels  $A_i$ , each of approximately the same size,  $\frac{1}{n_0}$ , and centered at  $z_i$ . If we use the approximation

$$\Lambda(A_i) = \int_{A_i} e^{\alpha + \beta' x(z)} dz \quad (16)$$

$$\approx |A_i| e^{\alpha + \beta' x_i} \quad (17)$$

$$\approx \frac{1}{n_0} e^{\alpha + \beta' x_i} \quad (18)$$

then the IPP model implies that the counts in each neighborhood  $A_i$  are generated via the Poisson LLM

$$N(A_i) \sim \text{Poisson} \left( \frac{1}{n_0} e^{\alpha + \beta' x_i} \right) \quad (19)$$

The log-likelihood of this model is (up to an additive constant)

$$\ell(\alpha, \beta) = \sum_{y_i=0} N(A_i)(\alpha + \beta' x_i) - \frac{1}{n_0} \sum_{y_i=0} e^{\alpha + \beta' x_i} \quad (20)$$

Since  $x(z)$  is continuous,

$$\sum_{y_i=0} N(A_i)(\alpha + \beta' x_i) = \sum_{y_i=0} \sum_{\substack{y_k=1 \\ z_k \in A_i}} \alpha + \beta' x_i \quad (21)$$

$$\approx \sum_{y_i=0} \sum_{\substack{y_k=1 \\ z_k \in A_i}} \alpha + \beta' x_k \quad (22)$$

$$= \sum_{y_k=1} \alpha + \beta' x_k \quad (23)$$

Therefore, (13) is almost exactly the same as the Poisson LLM log-likelihood for this discretized model. The only difference between the two is that in (20) we have also discretized the location of each presence sample to match its nearest background sample.

We could indeed fit an IPP model in exactly this way, by simply deleting the features of the presence samples and recording only how many fall into each background sample's surrounding pixel. Approximating the model in this way provides a simple way of accessing the generalizability of GLM methods (see Section 4).

As we have seen, this is essentially akin to replacing  $x_i$  for each presence point with the  $x_i$  of its nearest background sample. As we will see in Section 4.2, we can do essentially the same thing without making this approximation if we use a weighted GLM method.

## 2.5 Identifiability and Observer Bias

Observer bias poses one of the most serious challenges to valid inference in presence-only studies. Scientifically, we are interested in the *occurrence process*



consisting of all specimens of the species of interest. However, our data set comes from the *observation process* consisting only of the occurrences observed and reported by people.

Assume that each occurrence is observed with probability  $s(z)$ , which may depend on features of the geographic location  $z$ . If observation is independent across occurrences, then the observation process is an IPP with intensity

$$\lambda_{\text{obs}}(z) = \lambda_{\text{occ}}(z)s(z) \quad (24)$$

A presence-only data set only directly reflects  $\lambda_{\text{obs}}$ . Absent any assumptions about  $s$  the underlying intensity  $\lambda_{\text{occ}}$  is not identifiable.

One (optimistic) assumption we could make about  $s$  is that it is an unknown constant. In that case, by estimating  $\lambda_{\text{obs}}(z)$  we are also estimating  $\lambda_{\text{occ}}(z)$  up to an unknown constant of proportionality  $s$ . Even in this optimistic scenario we can only estimate relative occurrence intensities, not absolute intensities.

A bit more realistic is the assumption that  $s$  is an unknown function of  $z$ , but that  $s$  and  $\lambda_{\text{occ}}$  are known to depend on  $z$  through two disjoint feature sets. For instance, we could model  $\lambda_{\text{occ}}$  and  $s$  as log-linear in features  $x_1(z)$  and  $x_2(z)$  respectively

$$\lambda_{\text{obs}}(z) = \lambda_{\text{occ}}(z)s(z) \quad (25)$$

$$= e^{\tilde{\alpha} + \tilde{\beta}'x_1(z)} e^{\gamma + \delta'x_2(z)} \quad (26)$$

Then the observation process follows the log-linear model  $\lambda_{\text{obs}} = e^{\alpha + \beta'x(z)}$  with  $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$  and  $\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$ . Note that  $\tilde{\alpha}$  and  $\tilde{\beta}$  are the quantities of primary scientific interest, whereas  $\alpha$  and  $\beta$  are the parameters governing the process we actually observe. Nevertheless,  $\beta = \beta_1$  is still identifiable from the data because  $\beta$  is.<sup>1</sup>

As  $n_0, n_1 \rightarrow \infty$ , our estimate  $\hat{\beta}_1$  converges to the true value of  $\tilde{\beta}$ , the slope coefficients of  $\lambda_{\text{occ}}$ . However,  $\hat{\alpha}$  will converge not to  $\tilde{\alpha}$  but rather to  $\tilde{\alpha} + \gamma$ . Without knowing  $\gamma$  we have no way of estimating  $\tilde{\alpha}$ . Similarly, if  $x_1$  and  $x_2$  had overlapping (or linearly dependent) coordinates we could not estimate  $\tilde{\beta}$  for those coordinates.

To be more concrete, suppose koala occurrence is known to depend only on elevation, and that observer bias is known to depend only on proximity to roads. Then, despite the obvious observer bias in Figure 1 we could still estimate what elevation koalas prefer. By contrast, we could never estimate from this data whether koalas tend to avoid roads.

Even in the most optimistic scenario, we can estimate  $\alpha = \tilde{\alpha} + \gamma$  but it provides no real information about  $\tilde{\alpha}$ . Indeed, we have seen that the only role  $\hat{\alpha}$  plays in estimation to make  $\lambda$  integrate to  $n_1$ .

The distinction between  $\beta$  and  $\tilde{\beta}$  is important, but for most of this paper we will focus on estimation of  $\beta$ , the slope parameters of the process we get to observe. We revisit this distinction in Section 5.

<sup>1</sup>As with any regression adjustment scheme, we should proceed with caution here. If our linear model is misspecified (perhaps we should have included  $x_2^2$ ) and  $x_1$  is correlated with the missing variables, even regression adjustment will not remove all bias. In perverse situations it could even make the situation worse. Of course, this must be weighed against the fact that if there is observer bias, not accounting for it at all gives biased estimates too. See Section 5 for another option for dealing with observer bias.

## 2.6 Geographic Space and Feature Space

In the context of logistic regression, it will be more natural to think of the  $x_i$  being a sample of points in “feature space” (i.e. the range of  $x$ ) rather than as the features corresponding to a sample in the geographic domain  $\mathcal{D}$ . There is no real distinction between these two viewpoints, so long as we adjust for the fact that some values of  $x$  are more common in  $\mathcal{D}$  than others.

Suppose the  $z_i$  for presence samples ( $y_i = 1$ ) arise from an IPP with intensity  $\lambda(x(z))$ . We will show that the corresponding  $x_i$  are then an IPP with intensity  $\lambda_x(x) = \lambda(x)h(x)$ , where

$$h(x) = \int_{\{z: x(z)=x\}} dz \quad (27)$$

Suppose  $x$  were discrete. Then  $h(x)$  would be the total area of land with features equal to  $x$ , and if a presence sample were taken uniformly at random from  $\mathcal{D}$  ( $\beta = 0$ ) its probability of having features  $x$  would be proportional to that area. For continuous  $x$ ,  $h(x)$  is the marginal density of  $x$  in our domain  $\mathcal{D}$ , but the same intuition applies.

Suppose  $B$  is some subset of feature space, and consider the number  $N_x(B)$  of  $x_i$  falling in the set  $B$ . This is the same as the number of  $z_i$  falling in the inverse image  $A = x^{-1}(B) = \{z : x(z) \in B\}$ . That is,

$$N_x(B) = N(A) \sim \text{Poisson}(\Lambda(A)) \quad (28)$$

But

$$\Lambda(A) = \int_A e^{\lambda(x(z))} dz \quad (29)$$

$$= \int_B \int_{\{z: x(z)=x\}} e^{\lambda(x)} dz dx \quad (30)$$

$$= \int_B e^{\lambda(x)} h(x) dx \quad (31)$$

Furthermore, if  $B_1$  and  $B_2$  are disjoint sets, then so are  $A_1 = x^{-1}(B_1)$  and  $A_2 = x^{-1}(B_2)$ . It follows that  $N_x(B_1) = N(A_1)$  and  $N_x(B_2) = N(A_2)$  are independent, so the  $x_i$  satisfy our second definition of an IPP.

In terms of our “random sample” view of an IPP, the above derivation implies that  $\lambda_x$  integrates over the whole of feature space to  $\Lambda(\mathcal{D})$ , and the  $x_i$  corresponding to  $y_i = 1$  are distributed with density  $e^{\alpha+\beta'x}h(x)/\Lambda(\mathcal{D})$ .

## 3 Maximum Entropy / Conditional IPP

Another popular approach to modeling presence-only data is the Maxent method, proposed by Phillips et al. (2004). The authors begin by assuming that the presence samples  $z_1, \dots, z_{n_1}$  are a simple random sample from some probability distribution  $p(z)$ .

The authors adopt the view, inspired by information theory, that the estimate  $\hat{p}$  should have large entropy  $H(p) = -\int_{\mathcal{D}} p(z) \log(p(z)) dz$ , while also matching certain moments of the sample. Intuitively, the goal is for the estimate to be as “close to uniform” as possible, while still satisfying certain constraints

that make it resemble the empirical distribution. Indeed, if we maximized entropy with no constraints, we would estimate  $p$  as the uniform distribution over  $\mathcal{D}$ .

They propose to choose the  $p$  which maximizes  $H(p)$  subject to the constraint that the expectation of the features  $x(z)$  under  $\hat{p}$  matches the sample means of those features, i.e.

$$\frac{1}{n_1} \sum_{y_i=1} x_i = \int_{\mathcal{D}} x(z) \hat{p}(z) dz = \mathbb{E}_p x(z) \quad (32)$$

Phillips et al. (2004) show that this criterion is equivalent to maximizing the likelihood of the parametric model

$$p(z) = \frac{e^{\beta' x(z)}}{\int_{\mathcal{D}} e^{\beta' x(u)} du} \quad (33)$$

This is exactly the parametric form of  $p_\lambda$  for our log-linear IPP, and its log is exactly the partially maximized log-likelihood  $\ell^*(\beta)$ . The likelihood (3) is simply the likelihood of a simple random sample from  $p_\lambda$ , i.e. a conditional IPP. Indeed, the constraint (32) is nothing more than the score criterion for  $\beta$  in an IPP. This result may be found in Appendix A of Aarts et al. (2011).

The popular software package Maxent implements a method slightly more complex than the one originally proposed in 2004. First, it automatically generates a large basis expansion of the original features into many derived features (quadratic terms, interactions, step functions, and hinge functions of the original features). Then, it fits a model by optimizing an  $\ell_1$ -regularized version of the conditional IPP likelihood:

$$\sum_{y_i=1} \beta' x_i - n_1 \log \left( \int_{\mathcal{D}} e^{\beta' x(z)} dz \right) - \sum_j r_j |\beta_j| \quad (34)$$

The regularization parameters  $r_j$  are chosen automatically according to rules based on an empirical study of numerous presence-only data sets [6].<sup>2</sup>

Mathematically, the basis expansion only increases the length of the feature vector  $x(z)$ . Moreover, the  $\ell_1$  regularization scheme does not constitute an essential difference with the other methods considered here. One could (and often should) regularize the parameters of a fitted IPP process as well, especially if  $x(z)$  contains many features resulting from a large basis expansion.

Applying a penalty  $J(\beta)$  to the Maxent log-likelihood does not change the equivalence between the two models. Indeed, if we add a penalty term  $J(\beta)$  to the IPP log-likelihood (6), we still obtain (7) after differentiating with respect to  $\alpha$ . But then, when we partially maximize  $\ell(\alpha, \beta) - J(\beta)$  we simply obtain  $\ell^*(\beta) - J(\beta)$ , the penalized Maxent log-likelihood. Note that this equivalence depends on our not penalizing  $\alpha$ .

This argument generalizes immediately to a generic penalized likelihood method with a parametric form for  $\log \lambda(z)$ . We have established the following general proposition:

---

<sup>2</sup>In the notation of the Maxent papers the identities of  $\lambda$  and  $\beta$  are interchanged relative to the notation in this article.

**Proposition 1.** *Given some parametric family of real-valued functions  $\{f_\theta : \theta \in \mathbb{R}^p\}$  with penalty function  $J(\theta)$ , consider the penalized negative log-likelihood for an IPP with intensity  $e^{\alpha+f_\theta(x(z))}$*

$$g_1(\alpha, \theta) = - \sum_{y_i=1} \alpha + f_\theta(x_i) + \int_{\mathcal{D}} e^{\alpha+f_\theta(x(z))} dz + J(\theta) \quad (35)$$

*and the penalized negative log-likelihood for a conditional IPP with density proportional to  $e^{f_\theta(x(z))}$*

$$g_2(\theta) = - \sum_{y_i=1} f_\theta(x_i) + n_1 \log \left( \int_{\mathcal{D}} e^{\alpha+f_\theta(x(z))} dz \right) + J(\theta) \quad (36)$$

*Then (35) and (36) are equivalent in the sense if  $(\alpha, \theta)$  minimizes  $g_1$ ,  $\theta$  minimizes  $g_2$ , and if  $\theta$  minimizes  $g_2$ , there exists a unique  $\alpha$  for which  $(\alpha, \theta)$  minimizes  $g_1$ .*

*The same applies if we replace the integrals in (35) and (36) with sums over the background sample.*

*Proof.* Partially optimize  $g_1$  over  $\alpha$  as in (8) to obtain  $g_2$ .  $\square$

Thus we see that, while Maxent and the IPP appear to be different models with different motivations, they fit the exact same density  $p_\lambda$ . Fundamentally, this is a consequence of what we observed in Section 2.2: Maxent solves the same density estimation problem as step 1 of the IPP-fitting procedure, then skips step 2.

## 4 Logistic Regression

Another ostensibly different model for presence-only data is the so-called “naive” logistic regression. This approach treats presence-only modeling as a problem of classifying points as presence ( $y = 1$ ) or background ( $y = 0$ ) on the basis of their features. The logistic regression model treats  $n_1$ ,  $n_0$ , and the  $x_i$  as fixed and the  $y_i$  as random with

$$\mathbb{P}(y_i = 1|x_i) = \frac{e^{\eta+\beta'x_i}}{1 + e^{\eta+\beta'x_i}} \quad (37)$$

Superficially this approach may appear ad hoc and scientifically unmotivated compared to IPP or Maxent. Weighed against this concern is the fact that logistic regression is an extremely mature method in statistics, enjoying myriad well-understood and already-implemented extensions such as GAM, MARS, LASSO, boosted regression trees, and more.

Logistic regression modeling of presence-only data has often been motivated by analogy to logistic regression for presence-absence data. Since it is not known whether the species is present at or near the background examples, these are sometimes referred to as “pseudo-absences,” and the supposed naivete of the method refers to the fact that it treats them as actual absences. Various authors have introduced latent variables coding “true” presence or absence and have tried to fit this model via the EM algorithm or hierarchical methods [11], [10].

This interpretation raises once again the troublesome question of what it would actually mean for one of our randomly sampled background points to be a “true absence” or “true presence.” Would there need to be a specimen sitting directly on the location? Or is it enough for it to be within 100 m? 1 km? In fact, if we adopt the first view it would seem that our worries are over since we can assume there is a near-zero probability that a random geographic point coincides with the species of interest.

Fortunately we can sidestep these concerns entirely, since there are deep connections between the logistic regression and IPP models which yield a more straightforward interpretation. Warton & Shepherd (2010) showed that when the IPP model holds, so does the logistic regression model, and that if  $n_1$  is held fixed and  $n_0 \rightarrow \infty$ , the difference between the fitted  $\hat{\beta}$  for the two models converges to 0.

We will show that the log-linear IPP model implies that  $\mathbb{P}(y_i|x_i)$  is exactly as in (37), with the same slope parameters  $\beta$ . However, if this model is misspecified and  $n_1$  grows along with  $n_0$ , the two models’ estimates for  $\beta$  generally do not converge to the same limit. The limiting parameters of the logistic regression in general will depend on the limiting ratio of  $n_0$  to  $n_1$  (Fithian and Hastie, 2012).

Nevertheless, by using a weighted form of logistic regression we will see that we can exactly recover the IPP estimate for  $\hat{\beta}$  in finite samples. This implies in particular that packages implementing extensions of weighted logistic regression can be used to fit analogous extensions of the IPP model.

## 4.1 Case-Control Sampling

If we begin with a log-linear IPP model and condition on the size of our presence and background samples, the  $x_i$  are a mixture of two simple random samples, one from the density  $e^{\alpha+\beta'x}h(x)/\Lambda(\mathcal{D})$  and the other with density  $h(x)$ . It follows that

$$\mathbb{P}(y_i = 1|x_i, n_1) = \frac{n_1 e^{\alpha+\beta'x_i} h(x_i) / \Lambda(\mathcal{D})}{n_0 h(x_i) + n_1 e^{\alpha+\beta'x_i} h(x_i) / \Lambda(\mathcal{D})} \quad (38)$$

$$= \frac{e^{\eta+\beta'x_i}}{1 + e^{\eta+\beta'x_i}} \quad (39)$$

with  $e^\eta = \frac{n_1 e^\alpha}{n_0 \Lambda(\mathcal{D})}$ . This logic is depicted in Figure 3.

Thus, the log-linear IPP model implies the individual  $y_i|x_i$  follow a logistic regression model with the same slope parameters  $\beta$ .<sup>3</sup>

Thus, given any finite sample of presence and background points, we could either maximize the numerical IPP likelihood or the logistic regression likelihood, and in either case we would be fitting the same model. This fact alone does not guarantee we will obtain the same estimates  $\hat{\beta}$  in any given finite sample, but if the log-linear model is correctly specified then maximizing either likelihood gives a consistent estimator of the true  $\beta$ .

However, when the log-linear model is misspecified, the fitted slopes for logistic regression and numerical IPP will in general not converge to the same

---

<sup>3</sup>The  $y_i$  are technically not independent given  $n_0$  and  $n_1$  (if we knew the other  $n_1 + n_0 - 1$  labels, we would know the last as well). This is always true in case-control studies, but it is typically ignored since the dependence is weak for large samples.

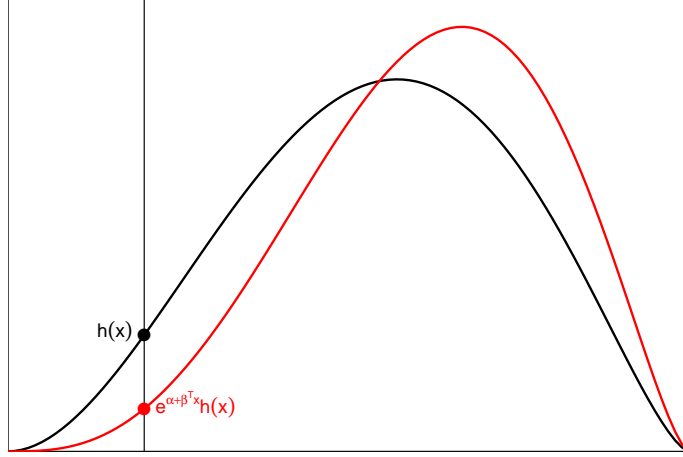


Figure 3: Presence-only sampling as case-control logistic regression.

limiting  $\beta$  if  $n_1$  and  $n_0$  grow large together. In fact, the limiting logistic regression parameters depend on the limiting ratio of  $n_1/n_0$ .

The reason is that when a model is misspecified (as are most parametric models, such as the linear model), then what we are estimating is the best parametric (linear) approximation in the population to the true function. But when we change the mix  $n_1/n_0$ , we are in effect changing the true population, and hence the parametric (linear) approximation (Fithian and Hastie, 2012).

If we modify the logistic regression procedure a bit, we need not wait for an infinite number of background samples. We can recover the same  $\hat{\beta}$  that we would estimate with a IPP using the same  $n_0$  background points.

## 4.2 Weighted Logistic Regression

Typically, although we use a finite background sample, we actually have an infinite (or at least much larger) reservoir of possible background points we could have used. Suppose we view each background point in our sample as a representative of many more background points which we only excluded for the purpose of computational convenience. To reflect this we might assign case weights to the samples

$$w_i = \begin{cases} W & y_i = 0 \\ 1 & \text{otherwise} \end{cases} \quad (40)$$

for some large number  $W$ . We would then obtain the weighted log-likelihood function

$$\ell_{\text{WLR}}(\eta, \beta) = \sum_i w_i \left[ y_i(\eta + \beta' x_i) - \log(1 + e^{\eta + \beta' x_i}) \right] \quad (41)$$

$$= \sum_{y_i=1} \eta + \beta' x_i - W \sum_{y_i=0} \log(1 + e^{\eta + \beta' x_i}) - \sum_{y_i=1} \log(1 + e^{\eta + \beta' x_i}) \quad (42)$$

If a unique MLE  $(\hat{\alpha}_{\text{IPP}}, \hat{\beta}_{\text{IPP}})$  exists for the IPP model, and  $(\hat{\alpha}_W, \hat{\beta}_W)$  solve (41) for weighting factor  $W$ , we have

$$\lim_{W \rightarrow \infty} \hat{\beta}_W = \hat{\beta}_{\text{IPP}} \quad (43)$$

We prove a more general version of this fact, again allowing the possibility of a general penalized likelihood approach.

**Proposition 2.** *Consider a family  $\{f_\theta : \theta \in \mathbb{R}^p\}$  where  $\theta \mapsto f_\theta(x)$  is concave and differentiable, with convex penalty function  $J(\theta)$ .*

*Suppose the penalized numerical negative log-likelihood for an IPP with intensity  $e^{\alpha+f_\theta(x(z))}$*

$$g_1(\alpha, \theta) = - \sum_{y_i=1} \alpha + f_\theta(x_i) + \frac{1}{n_0} \sum_{y_i=0} e^{\alpha+f_\theta(x_i)} + J(\theta) \quad (44)$$

*has a unique minimizer  $(\hat{\alpha}_{\text{IPP}}, \hat{\theta}_{\text{IPP}})$ . Also, define the penalized weighted logistic regression log-likelihood*

$$\begin{aligned} g_W(\eta, \theta) = & - \sum_{y_i=1} \eta + f_\theta(x_i) + \sum_{y_i=1} \log \left( 1 + e^{\eta+f_\theta(x_i)} \right) \\ & + W \sum_{y_i=0} \log \left( 1 + e^{\eta+f_\theta(x_i)} \right) + J(\theta) \end{aligned} \quad (45)$$

*Then if  $(\hat{\eta}_{W_k}, \hat{\theta}_{W_k})$  is any sequence of minimizers of  $g_{W_k}$  with  $W_k \rightarrow \infty$ , we also have*

$$\hat{\theta}_{W_k} \rightarrow \hat{\theta}_{\text{IPP}} \quad (46)$$

*Proof.* Define

$$\tilde{g}_W(\alpha, \theta) = g_W(\alpha - \log W n_0, \theta) - n_1 \log W n_0 \quad (47)$$

$$\begin{aligned} &= - \sum_{y_i=1} \alpha + f_\theta(x_i) + W \sum_{y_i=0} \log \left( 1 + \frac{1}{W n_0} e^{\alpha+f_\theta(x_i)} \right) \\ &+ \sum_{y_i=1} \log \left( 1 + \frac{1}{W n_0} e^{\alpha+f_\theta(x_i)} \right) + J(\theta) \end{aligned} \quad (48)$$

a shifted version of  $g_W$ . Since  $(\eta_{W_k} + \log W n_0, \theta_{W_k})$  minimize  $\tilde{g}_{W_k}$ , it suffices to show that any sequence of minimizers of  $\tilde{g}_{W_k}$  also satisfy (46).

In particular, we will show that as  $W \rightarrow \infty$ ,  $\tilde{g}_W \rightarrow g_1$  uniformly on compact subsets of  $\mathbb{R}^{p+1}$ . Since  $\tilde{g}_W$  and  $g_1$  are convex and the minimizer of  $g_1$  is unique, uniform convergence in any neighborhood of  $(\hat{\alpha}_{\text{IPP}}, \hat{\theta}_{\text{IPP}})$  is sufficient to prove the claim.

The difference between the two criteria is

$$\begin{aligned} \tilde{g}_W(\alpha, \theta) - g_1(\alpha, \theta) = & \sum_{y_i=0} \left( W \log \left( 1 + \frac{1}{W n_0} e^{\alpha+f_\theta(x_i)} \right) - \frac{1}{n_0} e^{\alpha+f_\theta(x_i)} \right) \\ & + \sum_{y_i=1} \log \left( 1 + \frac{1}{W n_0} e^{\alpha+f_\theta(x_i)} \right) \end{aligned} \quad (49)$$

For any compact  $\Theta \subseteq \mathbb{R}^{p+1}$ , we have

$$\sup_{\substack{1 \leq i \leq n_0 + n_1 \\ (\alpha, \theta) \in \Theta}} \alpha + f_\theta(x_i) = B < \infty \quad (50)$$

so that  $\frac{1}{Wn_0} e^{\alpha + f_\theta(x_i)}$  tends uniformly to 0 for all values of  $\alpha, \theta$  and  $x_i$  under consideration.

Using the Taylor expansion  $\log(1 + u) = u + O(u^2)$ , we obtain

$$\sup_{(\alpha, \theta) \in \Theta} |g_1(\alpha, \theta) - \tilde{g}_W(\alpha, \theta)| = O(W^{-1}) \quad (51)$$

proving the result.  $\square$

The above proof goes through with virtually no modification if we substitute for the logistic regression log-likelihood the poisson log-linear model log-likelihood:

$$\ell_{\text{WLLM}}(\eta, \beta) = \sum_i w_i \left[ y_i(\eta + \beta' x_i) - e^{\eta + \beta' x_i} \right] \quad (52)$$

$$= \sum_{y_i=1} \eta + \beta' x_i - W \sum_{y_i=0} e^{\eta + \beta' x_i} - \sum_{y_i=1} e^{\eta + \beta' x_i} \quad (53)$$

**Proposition 3.** *Under the same conditions as Proposition 2, if instead*

$$\begin{aligned} g_W(\eta, \theta) &= - \sum_{y_i=1} \eta + f_\theta(x_i) + \sum_{y_i=1} e^{\eta + f_\theta(x_i)} \\ &\quad + W \sum_{y_i=0} e^{\eta + f_\theta(x_i)} + J(\theta) \end{aligned} \quad (54)$$

then (46) holds as before for any sequence of minimizers of  $g_{W_k}$  with  $W_k \rightarrow \infty$ .

*Proof.* As before, define

$$\tilde{g}_W(\alpha, \theta) = g_W(\alpha - \log Wn_0, \theta) - n_1 \log Wn_0 \quad (55)$$

$$(56)$$

Then

$$\tilde{g}_W(\alpha, \theta) - g_1(\alpha, \theta) = \sum_{y_i=0} \left( \frac{1}{n_0} e^{\alpha + f_\theta(x_i)} - \frac{1}{n_0} e^{\alpha + f_\theta(x_i)} \right) + \sum_{y_i=1} \frac{1}{Wn_0} e^{\alpha + f_\theta(x_i)} \quad (57)$$

$$= \sum_{y_i=1} \frac{1}{Wn_0} e^{\alpha + f_\theta(x_i)} \quad (58)$$

which tends uniformly to zero on compact subsets of  $\mathbb{R}^{p+1}$ . The rest of the proof is the same.  $\square$

These two results also imply that logistic regression and poisson regression converge to each other when we upweight the negative examples. This phenomenon has a simple heuristic explanation. As we upweight the negative examples we drive all the fitted means toward zero, by driving  $\hat{\eta}$  to  $-\infty$ . There is hardly any difference between a  $\text{Poisson}(e^\lambda)$  random variable and a Bernoulli  $\left(\frac{e^\lambda}{1+e^\lambda}\right)$  random variable for very negative  $\lambda$ , and for that reason there is hardly any difference between the two GLMs.



### 4.3 Logistic Regression as Density Estimation

One interpretation of the results we have just reviewed is that in the context of presence-only data, logistic regression estimates the same density estimation problem as Maxent and the IPP do. Moreover, weighted logistic regression even finds the exact same estimates as the numerical IPP and Maxent procedures do.

Using logistic regression for density estimation is not without precedent. It was previously discussed in Section 14.2.5 of Hastie et al. (2009) as a means for turning the unsupervised problem of density estimation into a well-understood supervised classification problem of samples against background. The specific proposal in that book chooses a different weighting scheme (assigning half the total weight to each sample), which does not coincide exactly with IPP.

We believe that viewing logistic regression as a density estimation procedure resolves many of the conceptual misunderstandings that originally led to its labeling as “naive.”

### 4.4 Simulation Study: Weighted vs Unweighted Logistic Regression

We have seen that both weighted logistic regression (a.k.a. numerical IPP) and unweighted logistic regression estimate the same  $\beta$  parameter of the same IPP model, and when the background sample is much larger than the presence sample, the estimates  $\hat{\beta}$  are close to each other.

However, the weighted logistic regression estimate can converge much faster to the large-background-sample limit if the linear model is misspecified. We illustrate this phenomenon with a simulation study. We first generate  $n_1 = 1000$  positive examples from the normal distribution with mean 1 and standard deviation 2. Fixing the presence sample, we then repeatedly generate random background samples of various sizes  $n_0$ . The negative  $x_i$  are generated from the normal distribution with mean 0 and standard deviation 1.

For each background sample, we fit both a weighted ( $W = 1000$ ) and unweighted logistic regression to the combination of presence and background points. There are twenty replicates per value of  $n_0$ , and the results are shown in Figure 4. For relatively large sizes of background sample, there is very little sampling variability, but the logistic regression estimates carry a large bias that depends greatly on the size of the background sample. The limiting  $\hat{\beta}$ , to which both methods would converge given an infinite background sample, is depicted with a horizontal line.

Since the choice of background sample size is primarily a matter of convenience, it is preferable to use an estimator that depends on it as little as possible. When the linear model is misspecified (which is nearly always the case), we recommend the weighted logistic regression / IPP for this reason.

## 5 Pooling Data from Multiple Sources

We have seen that the IPP model is a unifying framework for understanding several popular approaches to modeling presence-only data. Its simple form induces familiar parametric forms for other quantities we might model as well, for example presence-absence and species count data.

**Weighted and Unweighted Estimates for Logistic Regression**

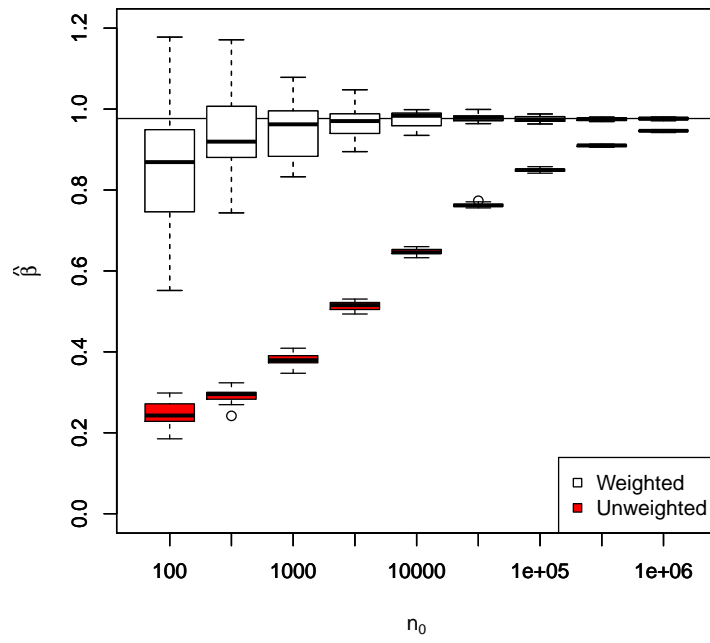


Figure 4: Unweighted logistic regression may require a very large background sample before convergence

If a team of surveyors exhaustively survey plots of land  $A_i$  with areas  $|A_i|$  and covariates  $x_i$ , then the number  $N_i$  of specimens they encounter will be distributed poisson with mean  $\Lambda(A_i) = |A_i|e^{\tilde{\alpha} + \tilde{\beta}'x_i}$ . Recall that  $\tilde{\alpha}$  and  $\tilde{\beta}$  are the parameters of the underlying occurrence process —  $\lambda_{\text{occ}}$  in (25). This is a poisson LLM with offsets  $\log |A_i|$ .<sup>4</sup>

If instead we only record whether or not at least one specimen was encountered, then we have a bernoulli GLM with complementary log-log link and offset  $|A_i|$

$$\mathbb{P}(N_i > 0 | x_i) = 1 - e^{-|A_i|e^{\tilde{\alpha} + \tilde{\beta}'x_i}} \quad (59)$$

The fact that one model yields coherent likelihoods for all these sampling schemes means we can pool together presence-only, presence-absence, and count data into a single log-likelihood.

Why might we want to do this? First, suppose we make the assumption of no selection bias, i.e.  $\beta = \tilde{\beta}$ . Then if we had access to a small presence-absence data set and a large presence-only data set, the presence-only data would help us to pin down  $\tilde{\beta}$  allowing the presence-absence data to more efficiently estimate  $\tilde{\alpha}$ .

In practice, it may be too much to assume the surveyors see every animal, but we could assume that at least there is no dependence on  $z$ , and surveyors see each animal independently with equal probability. In this case, the intercept for presence-absence data is not  $\tilde{\alpha}$  but  $\tilde{\alpha} - \varepsilon$ , where  $\varepsilon > 0$ . Presence-only sampling will not tell us the overall level of abundance but it might still give a useful lower bound if  $\varepsilon$  is not too large.

## 5.1 Estimating Sampling Bias

Suppose that we are not willing to assume away sampling bias, but we do believe that it is the same for several species. For instance, proximity to roads and cities may introduce a comparable observer bias on similar species.

With this assumption as motivation, Phillips et al. (2009) proposed using other species' sightings as background observations instead of randomly sampled locations. This method, called the “target-group background” (henceforth TGB) method by the authors, effectively “controls away” any observer bias that affects all species equally.

Unfortunately, the TGB approach also controls away any real environmental factor that affects overall abundance. If half of our environment is a lush rainforest and the other half is a desert wasteland, this method implicitly assumes that the overall abundance in each region is the same, and the only reason we have fewer samples from the desert is observer bias.

If we have several similarly-biased species with comparable observer bias, we might use the model

$$\lambda_{\text{occ},j}(z) = e^{\tilde{\alpha}_j + \tilde{\beta}_j'x(z)} \quad (60)$$

$$\lambda_{\text{obs},j}(z) = e^{\tilde{\alpha}_j + \gamma + (\tilde{\beta}_j + \delta)'x(z)} \quad (61)$$

where  $j$  indexes species.

All the parameters of this model are identifiable once we have

---

<sup>4</sup>We should proceed with caution in modeling count data as Poisson, since the actual counts are likely to be overdispersed.

1. Presence-only data from every species.
2. Presence-absence or count data for a single species (say,  $j = 1$ ).

The presence-only data lets us independently estimate  $\alpha_j$  and  $\beta_j$  for each species, while the presence-absence or count data lets us estimate  $\tilde{\alpha}_1$  and  $\tilde{\beta}_1$ . This is enough to estimate the other  $\tilde{\alpha}_j$  and  $\tilde{\beta}_j$  because

$$\tilde{\alpha}_j = \alpha_j - \gamma \quad (62)$$

$$= \alpha_j + \tilde{\alpha}_1 - \alpha_1 \quad (63)$$

$$\tilde{\beta}_j = \beta_j - \delta \quad (64)$$

$$= \beta_j + \tilde{\beta}_1 - \beta_1 \quad (65)$$

If there is imperfect but unbiased detection in the presence-absence study, then as before we can only estimate  $\tilde{\alpha}_j - \varepsilon$  and not  $\tilde{\alpha}_j$ . However, this would not affect identifiability for  $\tilde{\beta}_j$ .

Depending on the species involved and the amount of data available, we may also benefit by shrinking estimates of the  $\alpha_j$  toward each other or sharing information in some other way.

## 5.2 Simulation Study

In this section we simulate a simple fictional landscape with ten species, two environmental features, and a third feature which induces observer bias.

Our geographic domain is the unit square  $[0, 1]^2$ , and the two environmental variables are identified with the geographic axes. For simplicity we have discretized the geographic space into pixels, so there is no distinction between the LLM and the IPP.

The first variable, “wasteland,” is negatively associated with all ten species by varying amounts, and the second, “elevation,” is positively associated with some species and negatively for others. The particular  $\tilde{\beta}_j$  for each species is randomly generated.

The region also contains four “towns” in which human population, the third variable, is large. The human population does not affect the species, but it does have a multiplicative effect on the rate at which sightings occur. Figure 5 displays the relative observance and occurrence intensities for two species.

Our data consist of presence-only data sets for each species, each with 300 points in expectation. First we simply fit an IPP to each species using the other species’ observations as a background sample, following the TGB proposal. This method essentially looks for contrasts between the observation intensities of the different species. Its estimates are listed in Table 1 under the heading TGB.

It succeeds in removing the observer bias from the estimated intensities, but it additionally removes the average effect of the wasteland on all species, producing severely biased estimates. Given the data we have and no other information or assumptions, there is fundamentally no way to distinguish between observer bias and a legitimate environmental effect common to the different species.

If we knew ahead of time that the human population variable only affects  $s(z)$  and the environmental variables only affect  $\lambda_{\text{obs}}$ , we could use the regression-adjustment approach of Section 2.5 for each species. Because that assumption is correct in this case, regression adjustment successfully adjusts for the observer

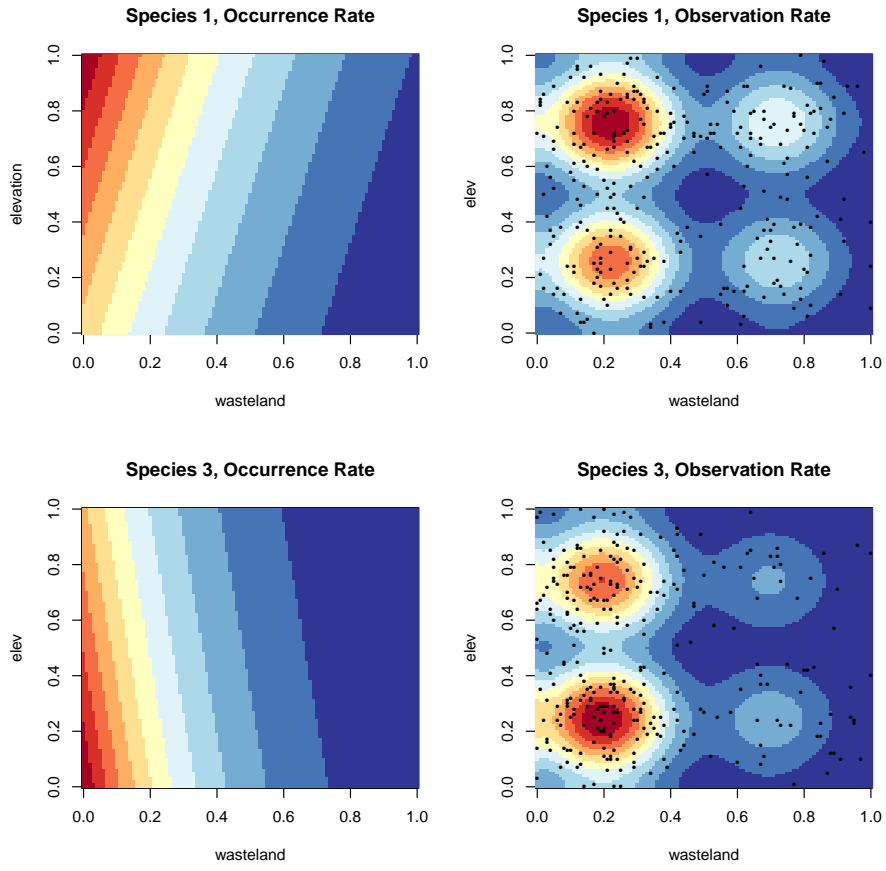


Figure 5:  $\lambda_{\text{occ}}$  and  $\lambda_{\text{obs}}$  for two species. Red is high, blue is low.

bias and not for the true environmental effect. The estimates from this method are in Table 1 under the heading RA (for regression adjustment).

If we were not willing to make this assumption (for instance, a priori it could be that human population has a real effect on the species' occurrence process), we could still account correctly for observer bias by fitting the model defined by (60-61). This model requires additional data: unbiased presence-absence or count data for at least one species. Here we independently simulate count data for species 1, "collected" from 300 randomly chosen sites. Since the distribution of the counts depends on sampling area  $|A_i|$  and detection probability  $\varepsilon$ , we choose these implicitly so that their means are proportional to  $e^{\hat{\beta}'_1 x_i}$  (equal sampling area) and the average site has expected count of 2.

Because we have discretized geographic space, the presence-only data can also be represented as counts over the pixels. This implies we can maximize likelihood for (60-61) by simply fitting a single poisson LLM to the concatenated records of the 11 (10 presence-only and 1 count) data sets, using the R function glm. The estimates from this method are in Table 1 under the heading DP (for data pooling).

Species	Log Human Pop.				Wasteland				Elevation			
	True	TGB	RA	DP	True	TGB	RA	DP	True	TGB	RA	DP
1	0.00	0.01	0.00	-0.08	-1.55	0.93	-1.23	-1.52	0.42	0.40	0.43	0.33
2	0.00	-0.17	0.00	-0.24	-2.09	-0.07	-2.16	-2.45	0.98	0.87	0.89	0.79
3	0.00	0.05	0.00	-0.07	-2.79	-0.72	-2.67	-2.96	-0.39	-0.44	-0.29	-0.38
4	0.00	-0.18	0.00	-0.26	-1.43	0.77	-1.31	-1.60	-1.04	-1.15	-1.03	-1.13
5	0.00	0.33	0.00	0.22	-1.96	0.15	-1.98	-2.27	1.78	1.52	1.51	1.42
6	0.00	-0.07	0.00	-0.15	-2.07	0.00	-1.92	-2.21	-2.31	-3.19	-2.90	-3.00
7	0.00	0.24	0.00	0.13	-2.35	-0.31	-2.34	-2.63	0.88	0.38	0.46	0.36
8	0.00	-0.01	0.00	-0.10	-1.88	-0.05	-2.11	-2.40	0.04	0.24	0.32	0.22
9	0.00	0.05	0.00	-0.06	-2.99	-1.12	-3.10	-3.39	1.01	0.85	0.90	0.81
10	0.00	0.01	0.00	-0.08	-1.93	0.09	-1.99	-2.28	0.43	0.43	0.50	0.41

Table 1: Results of three methods for correcting observer bias. The target-group background method succeeds in controlling for the Human Population variable, but in so doing controls away the Wasteland variable.

## 6 Discussion

We have shown here that the IPP, Maxent, and weighted logistic regression are equivalent in several senses:

1. All may be derived from the IPP model.
2. All may be thought of as performing the same parametric density estimation problem, which amounts to fitting  $\beta$ .
3. All fit the same  $\hat{\beta}$  given the same finite presence and background samples.

The only difference between the IPP and the other two methods is that it fits an intensity equal to  $n_1$  times the fitted density.

These findings remain the same if we replace the  $\beta'x_i$  term in the exponent with a smooth parametric model that is convex in its parameters  $\theta$ , or apply a convex penalty to  $\beta$ .

Logistic regression is one of the most widely applied methods in statistics. For decades, applied statisticians have been developing, studying, and using variations on logistic regression to solve classification problems in statistics. R packages exist for fitting generalized additive models (GAMs), boosted regression trees, MARS, and every manner of tailored regularization schemes (see, e.g., [3]).

All of these methods are well-understood within the context of logistic regression, and we believe that the most important practical implication of the equivalence between the IPP model and weighted logistic regression is that all of these methods can now be equally well-understood within the context of the IPP model.

For instance, we can fit an IPP / Maxent version of boosted regression trees with the following single line of R:

```
boosted.ipp <- gbm(y~., family="bernoulli", data=dat, weights=1E3^(1-y))
```

For an IPP / Maxent version of LASSO, ridge, or the elastic net:

```
lasso.ipp <- glmnet(dat.x, dat.y, family="binomial", weights=1E3^(1-y))
```

For an IPP GAM:

```
gam.ipp <- gam(y~s(x1)+s(x2)+x3*x4, family=binomial, weights=1E3^(1-y))
```

Finally, the same IPP modeling framework induces likelihoods for other forms of data which may not have the same biases as presence-only data sets, which may then be combined with the presence-only data into a single modeling function. We have demonstrated one technique for using this data to correct for observer bias.

This added flexibility promises to provide a powerful tool to modelers of presence-only data.

## References

- [1] G. Aarts, J. Fieberg, and J. Matthiopoulos. Comparative interpretation of count, presence-absence and point methods for species distribution models. *Methods in Ecology and Evolution*, 2011.
- [2] C. Gaetan and X. Guyon. *Spatial statistics and modeling*. Springer Verlag, 2009.
- [3] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*, volume 1. Springer Series in Statistics, 2009.
- [4] CR Margules, MP Austin, D. Mollison, and F. Smith. Biological models for monitoring species decline: The construction and use of data bases [and discussion]. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 344(1307):69–75, 1994.
- [5] S.J. Phillips, R.P. Anderson, and R.E. Schapire. Maximum entropy modeling of species geographic distributions. *Ecological modelling*, 190(3):231–259, 2006.

- [6] S.J. Phillips and M. Dudík. Modeling of species distributions with maxent: new extensions and a comprehensive evaluation. *Ecography*, 31(2):161–175, 2008.
- [7] S.J. Phillips, M. Dudík, J. Elith, C.H. Graham, A. Lehmann, J. Leathwick, and S. Ferrier. Sample selection bias and presence-only distribution models: implications for background and pseudo-absence data. *Ecological Applications*, 19(1):181–197, 2009.
- [8] S.J. Phillips, M. Dudík, and R.E. Schapire. A maximum entropy approach to species distribution modeling. In *Proceedings of the twenty-first international conference on Machine learning*, page 83. ACM, 2004.
- [9] J. Andrew Royle, James D. Nichols, and Marc Kry. Modelling occurrence and abundance of species when detection is imperfect. *Oikos*, 110(2):353–359, 2005.
- [10] J.A. Royle, R.B. Chandler, C. Yackulic, and J.D. Nichols. Likelihood analysis of species occurrence probability from presence-only data for modelling species distributions. *Methods in Ecology and Evolution*, 2012.
- [11] G. Ward, T. Hastie, S. Barry, J. Elith, and J.R. Leathwick. Presence-only data and the em algorithm. *Biometrics*, 65(2):554–563, 2009.
- [12] D.I. Warton and L.C. Shepherd. Poisson point process models solve the “pseudo-absence problem” for presence-only data in ecology. *The Annals of Applied Statistics*, 4(3):1383–1402, 2010.